Red-Teaming for Generative AI: Silver Bullet or Security Theater?

Year: 2024 | Citations: 102 | Authors: Michael Feffer, Anusha Sinha, Zachary Chase Lipton, Hoda Heidari

---

Abstract

In response to rising concerns surrounding the safety, security, and trustworthiness of Generative AI (GenAI) models, practitioners and regulators alike have pointed to AI red-teaming as a key component of their strategies for identifying and mitigating these risks. However, despite AI red-teaming's central role in policy discussions and corporate messaging, significant questions remain about what precisely it means, what role it can play in regulation, and how it relates to conventional red-teaming practices as originally conceived in the field of cybersecurity. In this work, we identify recent cases of red-teaming activities in the AI industry and conduct an extensive survey of relevant research literature to characterize the scope, structure, and criteria for AI red-teaming practices. Our analysis reveals that prior methods and practices of AI red-teaming diverge along several axes, including the purpose of the activity (which is often vague), the artifact under evaluation, the setting in which the activity is conducted (e.g., actors, resources, and methods), and the resulting decisions it informs (e.g., reporting, disclosure, and mitigation). In light of our findings, we argue that while red-teaming may be a valuable big-tent idea for characterizing GenAI harm mitigations, and that industry may effectively apply red-teaming and other strategies behind closed doors to safeguard AI, gestures towards red-teaming (based on public definitions) as a panacea for every possible risk verge on security theater. To move toward a more robust toolbox of evaluations for generative AI, we synthesize our recommendations into a question bank meant to guide and scaffold future AI red-teaming practices.